Machine Learning Applications in Treatment Pattern Analysis Using Real-World Data: A Scoping Review

Putnam
Inizio Advisory

Katarzyna Jabłońska [1], Adrianna Czubin[1], Maja Ludwikowska[1], Sylvaine Barbier[2], Renata Majewska[1]
[1] Putnam, Poland, [2] Putnam, France

MSR140

Introduction

- While clinical trials remain the gold standard for assessing treatment efficacy and safety, there is growing interest in studies of treatment patterns based on real-world data (RWD) from electronic health records, registries, and claims databases.
- Analysing treatment patterns in RWD is challenging due to inconsistent definitions, large datasets, incomplete data, and evolving clinical practices.
- Advanced approaches such as machine learning (ML), increasingly applied in RWD studies (particularly for survival prediction), can enhance treatment pattern analyses by handling data complexity, variability, and large-scale combinations.
- Evidence on applying ML to treatment pattern analysis seem to be still limited.

Objective & Methods

- We aimed to evaluate recent applications of ML techniques for treatment pattern analysis using RWD by conducting a scoping review of observational studies using large real-world datasets.
- Eligible studies used ML methods to analyse treatment patterns and were indexed in MEDLINE or Embase (via OVID) by the search date (21st May 2025), with no restrictions on publication year.
- From the included studies, we extracted information on applied ML methodologies. Key characteristics of ML algorithms were assessed, and methods were summarised and compared in terms of their robustness and type of use.

Results

- From 233 abstracts screened, we identified 16 eligible studies, published between 2016 and 2024.
- 9 studies used ML approaches to analyse treatment patterns [1-9], while the remaining 7 studies applied ML methods to build treatment lines or to predict treatment events such as switches or add-ons [10-16].
- The most frequently used methods were K-means clustering [n=4], time-sequence clustering [n=3], and ATLAS (Analysis of Treatment Lines using Alignment of Sequences) [n=3] (**Figure 1**). The remaining methods included tree-based prediction models [n=2], hierarchical clustering [n=2], SNF (Similarity Network Fusion) [n=1], and LDA (Latent Dirichlet Allocation) [n=1].
- A summary of ML algorithms identified in this scoping review, with their strengths and weaknesses, is presented in **Table 1**. Examples of the use of selected methods are presented in **Figure 2** and **Figure 3**.

Table 1. Comparison of the extracted ML algorithms

Method	Description	Strengths ✓	Limitations x
K-means or hierarchical clustering [2, 3, 6, 9, 11, 13]	Groups patients into clusters based on treatment similarity.	Simple, efficient, widely used.	Ignores treatment order; may oversimplify trajectories.
Time-sequence clustering [1, 4, 5]	Extends K-means by incorporating temporal order of treatments.	Captures sequential patterns; more realistic trajectories.	More complex and computationally demanding.
ATLAS [12, 15, 16]	Builds treatment lines based on RWD and theoretical schemes or cycles.	Clinically interpretable treatment lines.	Less scalable for large datasets.
Tree-based prediction models [10, 14]	Predicts treatment events (e.g., switches, add-ons) using decision trees / ensembles.	Flexible, strong predictive power, handles complex interactions.	Require labeled outcomes; risk of overfitting; less trajectory-focused.
SNF [7]	Merges multiple similarity networks (e.g., clinical, genomic, treatment).	Integrates multimodal data; reveals hidden patient subgroups.	Computationally intensive; harder to interpret clinically.
LDA [8]	Identifies latent “topics” representing co-occurring treatments.	Uncovers hidden structures; probabilistic, flexible.	Needs parameter tuning; topics may be hard to interpret.

Figure 1. ML algorithm frequencies

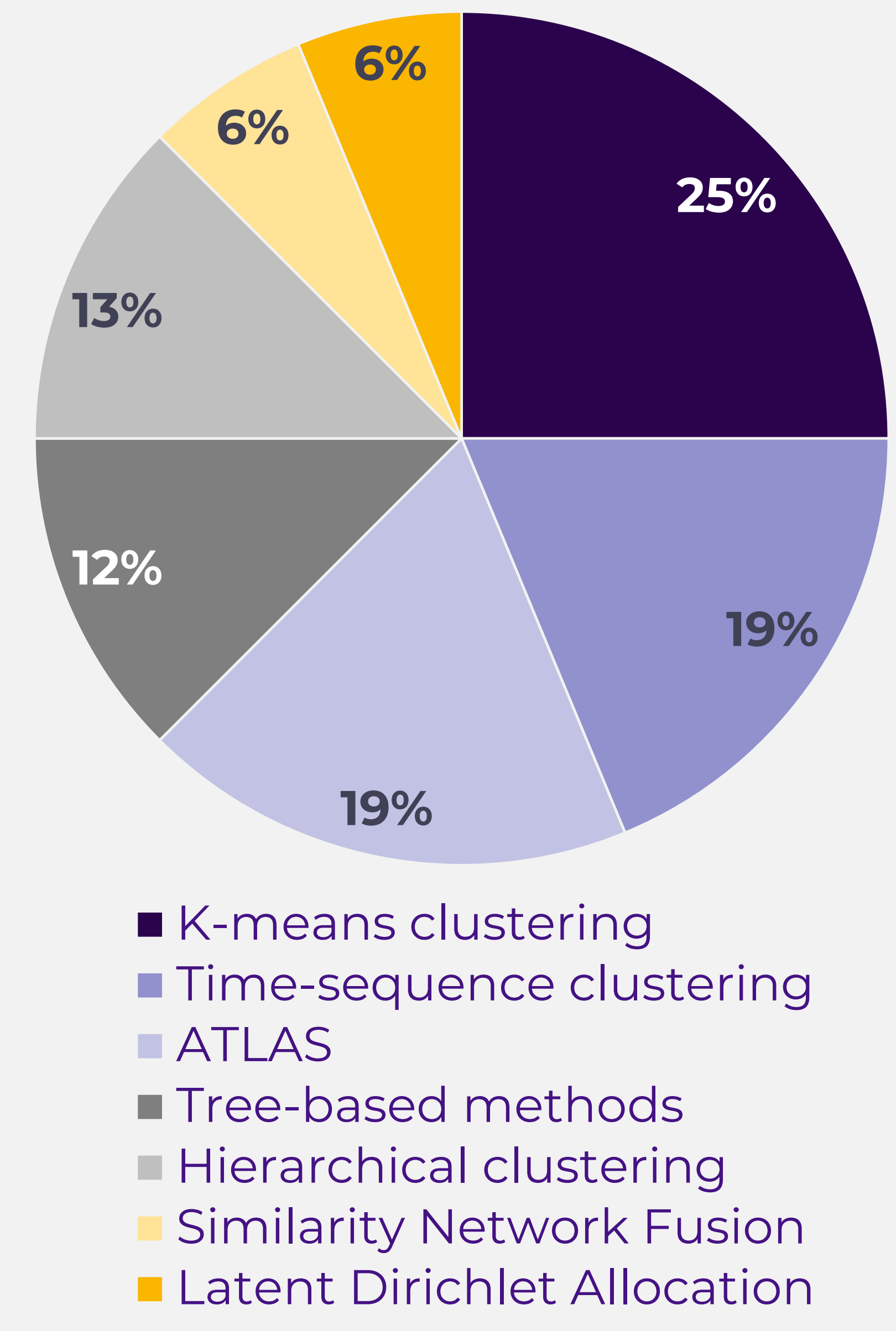


Figure 2. Example of using K-means clustering

Patients with similar treatment patterns are grouped into clusters, with colors indicating specific patterns. Reproduced from [2].

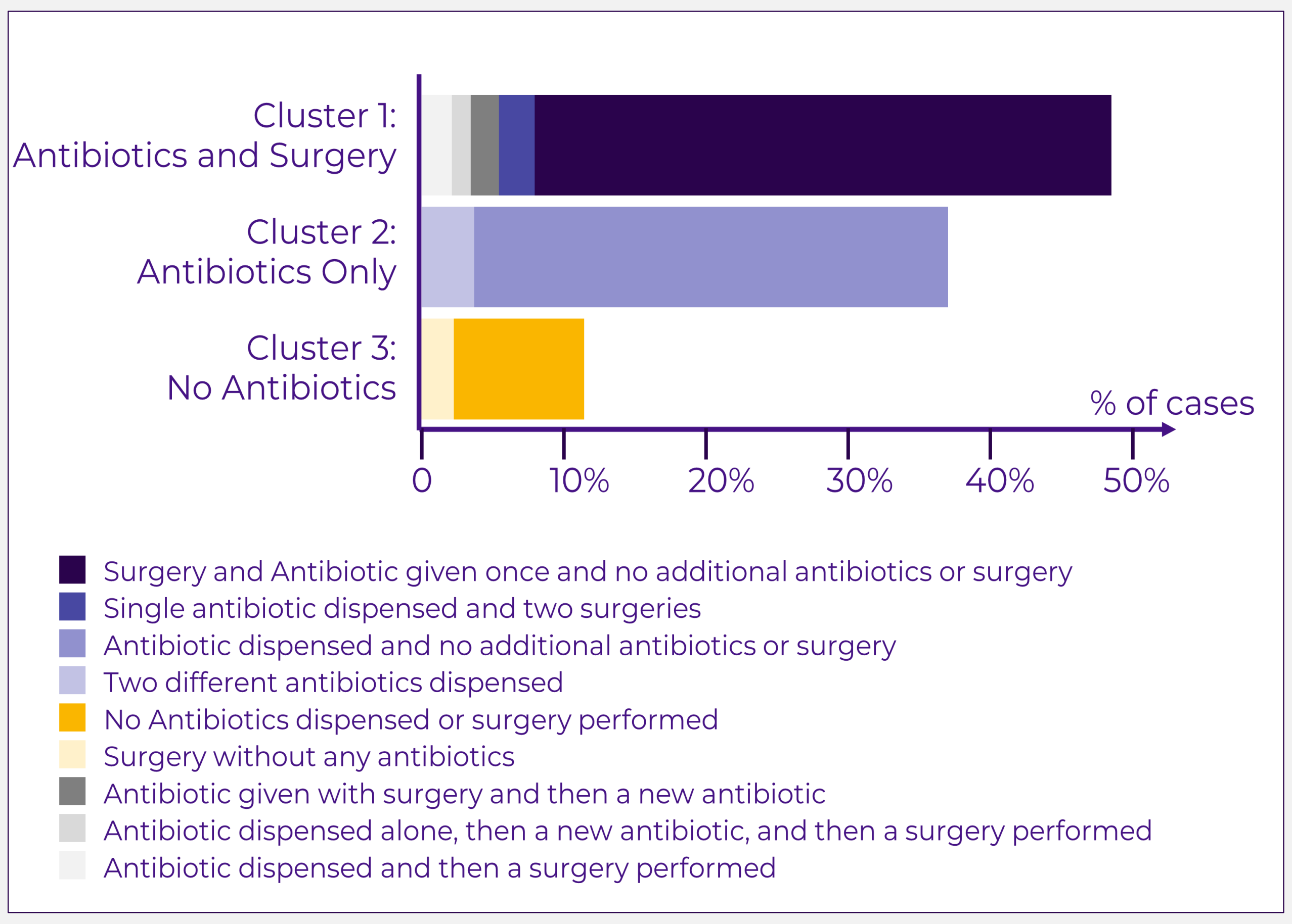
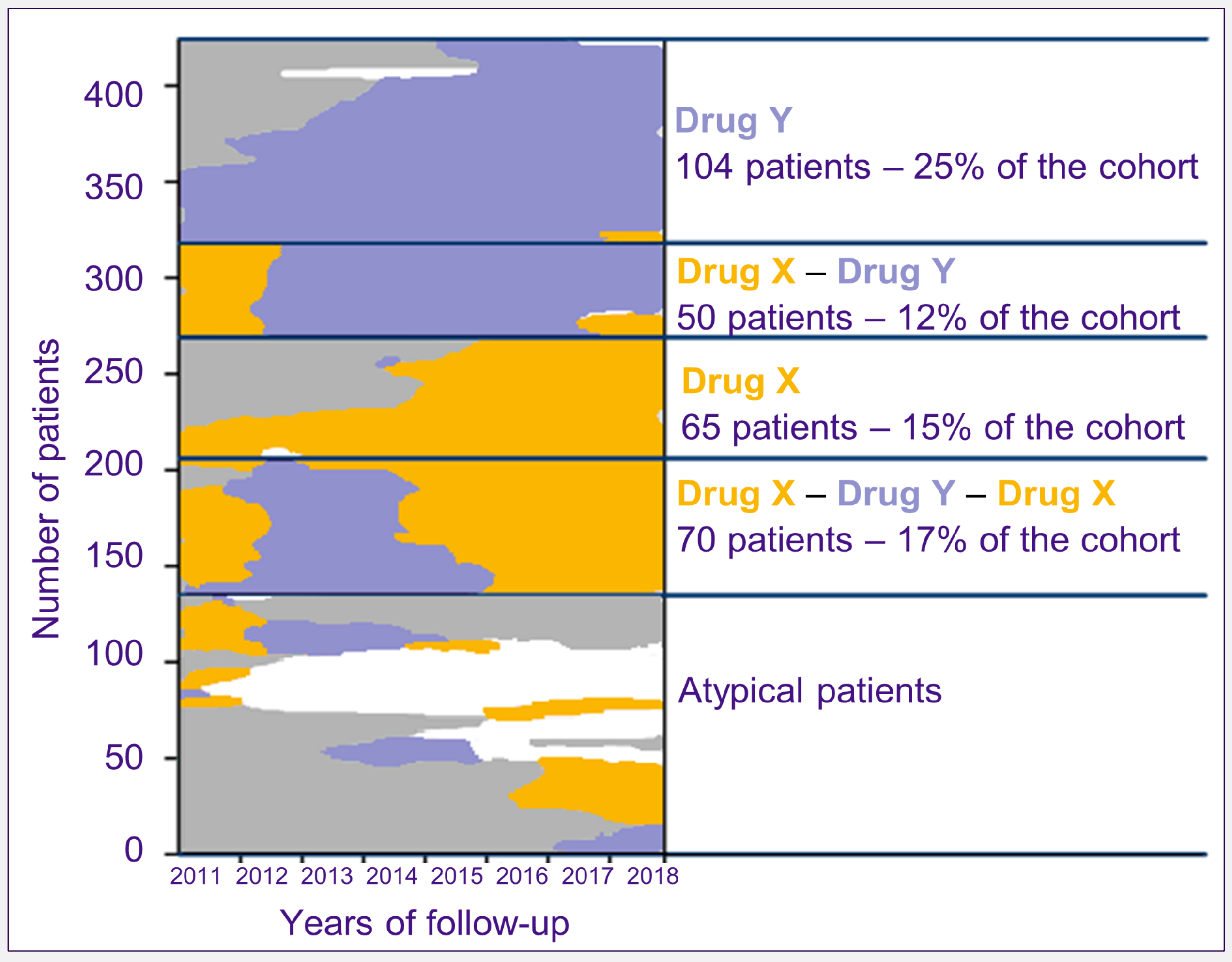


Figure 3. Example of using time-sequence clustering

Treatment at each follow-up time point and the treatment duration are clearly visualized. In-house case-study.



Discussion & Conclusion

- ML approaches can address key challenges in RWD treatment pattern analysis, such as data complexity and variability. The scope of their applications has expanded considerably, with the bulk of empirical evidence accumulating only within the past few years.
- Clustering approaches, particularly K-means, have been most widely adopted, effectively grouping patients based on similarities in treatment patterns. Time-sequence-based methods extend this approach by capturing how sequences evolve over time, while ATLAS focuses on building treatment trajectories, enabling the identification of common treatment lines across patients.
- Tree-based models have demonstrated strong potential for forecasting treatment events, supporting more patient-centered evaluations. More advanced methods (SNF, LDA), though less frequently used, highlight opportunities for uncovering hidden structures.
- However, many of these advanced approaches require substantial computational power and resources, which can pose practical challenges for large-scale implementation. Future research should broaden the use of advanced ML techniques in treatment pattern analysis, which hold promise for enhancing real-world evidence generation and complementing clinical trial data.

References

1. Chouaid, et al., 2022. 10.1200/CCI.21.00108
2. Hur, et al., 2024. https://dx.doi.org/10.1016/j.prevetmed.2023.106112
3. Rush, et al., 2024. https://dx.doi.org/10.1186/s12911-024-02469-4
4. Assie, et al., 2023. https://dx.doi.org/10.1016/j.resmer.2023.101051
5. Tredan, et al., 2022. https://dx.doi.org/10.1177/11769351221135134
6. Huan, et al., 2021. https://dx.doi.org/10.1155/2021/5590743
7. Chen, et al., 2020. https://dx.doi.org/10.1016/j.artmed.2019.101782
8. Fohner, et al., 2019. https://dx.doi.org/10.1093/jamia/ocz106
9. Kan, et al., 2016. https://dx.doi.org/10.1016/j.clinthera.2016.01.016
10. Breitenstein, et al., 2022. https://dx.doi.org/10.3389/fphar.2022.954393
11. Sessa, et al., 2021. https://dx.doi.org/10.1002/pds.5305
12. Perrot, et al., 2021. https://dx.doi.org/10.1016/S2152-2650%2821%2902090-5
13. Wang, et al., 2022. https://dx.doi.org/10.1016/j.jval.2023.03.1590
14. Touzeni, et al., 2019. https://dx.doi.org/10.1016/j.jval.2019.09.1537
15. Laurent, et al., 2020. https://static.hevaweb.com/web/PDF/ddc03b66d46f5-janssen-atlas-poster-afcro2020-v1r3-web.pdf
16. Prodel, et al., 2019. 10.1109/CIBCB.2019.8791467

Abbreviations

ATLAS: Analysis of Treatment Lines using Alignment of Sequences; LDA: Latent Dirichlet Allocation; ML: Machine Learning; RWD: Real-World Data; SNF: Similarity Network Fusion.

Contact
Sylvaine Barbier
Sylvaine.Barbier@putassoc.com

Find out more at putassoc.com